Data Understanding :

- Numeric Dataset - Attributes are of type int and float
- No of entries=>425262
- 10 columns have missing values (around 10 values were missing)
- Summary statistics=> SoilHeatFlux_DiffVolt1 column has extreme values like -9999 as the least=> requires cleaning
- Unique values present in each column=>SoilHeatFlux_DiffVolt2 is a column having all the entries as 0 => needs to be deleted
- Correlation heat matrix=>

```
The variable most positively correlated with Soil Heat Flux is: VolSWC_20cm (Correlation: 0.08)
The variable most negatively correlated with Soil Heat Flux is: NET_Radiation (Correlation: -0.82)
```

  => these columns are definitely a part of the features to be selected
- Histogram of soil heat flux=>skewed distribution=>outliers
- Box plot , hexbin(also tells about correlation)=> outliers
- Diurnal variation on an hourly basis=> soil heat flux - mostly negative - heat is being conducted downward- It tends to be most negative during the day, likely due to the heating of the surface by solar radiation. At night, the soil heat flux becomes less negative or even slightly positive, suggesting that heat is being conducted upward from the soil.
- Temporal variation - The soil heat flux remains relatively constant at a negative value most of the time, but it experiences sudden, sharp increases to positive values.
- Rolling mean and std deviation - The soil heat flux fluctuates periodically, with periods of high and low values. The average trend is downward, indicating a net heat loss from the soil.
- The ACF measures the correlation between a variable's current value and its past values- flat acf- white noise process, where the values are random and uncorrelated with each other.
- Analyzing how the heat flux and other parameters vary in a year using plots

Data Preprocessing:

- Dropping a column having zero as its value
- Renaming column for easy understanding
- Mean Imputation to fill the values
- Scaling the data using minmax scaler
- Removing outliers using IQR
- Feature selection using Univariate score,RFE Ranking, Random Forest Importance
- Univariate - statistical method (chi squared , ANOVA)
- RFE- Iteratively removes features that have the least impact on the model's performance.
- Random forest Importance- Measures the importance of each feature based on how much the model's prediction accuracy decreases when that feature is permuted.
- Recommended features: Volume 20cm, Volume 10cm, Net radiation, Conductivity, Soil Temperature 20cm ,Soil Temperature 10cm ,Soil Temperature 80cm
  DateTime or Month (from Date/Time field)